Signature Genes in Human Heart Failure Based on Gene Expression Analysis: Can We Identify a Unique Set?

Haiying Wang, Huiru Zheng, *Member, IEEE*

***Abstract*—Dilated Cardiomyopathy is one of leading courses of heart failure. Recent advances in microarray technology have promised significant advantages in understanding the molecular mechanisms underlying dilated cardiomyopathy and heart failure. Several microarray studies have successfully yielded a set of signature genes associated with heart failure. However, it has been found that the overlap of these heart failure associated genes derived from different experiments is very small. Based on the analysis of two publicly available microarray datasets associated with heart failure with three types of machine learning and statistical prediction models, this paper explores this phenomenon. We found that there is no unique set of genes associated with heart failure. Many sets of genes can achieve very high prediction accuracy. In order to identify biomarkers in human heart failure, it may not be sufficient to just focus a certain number of top genes. Such main candidates should be chosen from the much longer list of genes.**

I. Introduction

DILATED cardiomyopathy (DCM), a disorder of cardiac muscle, is a leading course of heart failure (HF) [1], [2]. In DCM, the heart muscle becomes weakened and one or both ventricles enlarge, making the heart pump blood less efficiently. The decreased heart function will eventually affect the working of other vital organs such as liver and lung. The causes of DCM are heterogeneous and despite many efforts, there is no specific genetic defect that has yet been well established.

Recent advances in microarray technology have promised significant advantages in the identification of biomarker associated with HF and a better understanding of the molecular mechanisms underlying DCM and HF. Several microarray studies have successfully yielded a set of HF signature genes [2-5]. For example, based on the comparative analysis of gene profiles of seven nonfailing heart and eight failing human heart with a diagnosis of end-stage DCM, Tan *et al.* [3] identified 103 differentially expressed genes which can be used to represent a gene fingerprint for human heart failure. They divided these genes into 10 groups: biomarker, myofibrillar, extracellular matrix/cytoskeletal, proteolysis stress, metabolism, apotosis/inflammatory, signal transduction, immune system and genes of unknown function. In an attempt to identify a common gene expression signature in DCM across multiple microarray studies, Barth *et al.* [2] performed 2 genome-wide expression studies using cDNA and short-oligonucleotide platforms which comprised independent septal and left ventricular tissue samples from 40 patient samples. Together with 2 publicly available datasets, they studied gene profiles associated with HF from a total of 108 myocardial samples and identify a robust set of 27 genes which can differentiate DCM samples from healthy subjects with an accuracy of 90%. Wittchen *et al.* [4] studied the genomic expression profile of inflammatory cardiomyopathy and identified two significantly altered gene networks centred around the cysteine-rich angiogenic inducer 61 (CYR61) and adiponectin (APN) gene. They argued that dysbalance between the CYR61 and APN networks could have a pathogenic role in inflammatory cardiomyopathy and may contain novel therapeutic targets. More recently, Camargo and Azuaje [5] integrated three publicly available microarray datasets. Differentially expressed genes were evaluated in the context of a global protein protein interaction network. The main outcome of this study is a set of integrated, potentially novel DCM signature genes including PICK1, DYNLL1, ODC1, HTRA1 and HMGN2.

However, it has been found that lists of biomarkers in HF derived from different experiment studies had only a few genes in common. For example, among 103 differentially expressed genes identified by Tan *et al.* [3] and 27 genes found by Barth *et al.* [2], they are only 4 genes appearing in both sets. Camargo and Azuaje [5] used prediction analysis of microarray (PAM) technique to identify 47, 3, and 36 significant class predictor genes (CP) whose expression profile showed strong discriminative capability between DCM and non-DCM samples from three heterogeneous, independent datasets respectively. Surprisingly, none was shared by these three sets. Such the lack of agreement between the sets of signature genes derived from different studies was also found in breast cancer studies [6] and other human diseases [7], [8].

This paper aims to explore this phenomenon in the context of HF. Based on the analysis of two publicly available datasets using machine learning and statistical prediction models, the following question are addressed: (1) how many genes do we need to build a prediction model for the classification of DCM samples; (2) Is there a unique set? and (3) how to explain the disparity of gene sets reported by independent studies. The remainder of the paper is organized as follows. Section II briefly describes the dataset

Manuscript received July 5, 2008. This work was supported in part by a grant from EU FP6, CARDIOWORKBENCH.

HY. Wang and H. Zheng are with the School of Computing and Mathematics, University of Ulster, Newtownabbey, Co. Antrim, BT37 0QB, U.K. (e-mail: hy.wang@ulster.ac.uk; h.zheng@ulster.ac.uk).

under study, followed by a description of the prediction models and statistical evaluation techniques. The results are presented in Section III. The discussion of results and conclusions, together with future research, are given in Section IV.

II. Methodology

A. Datasets under study

The two microarray datasets used in this paper were derived from a study on DCM published by Barth *et al.* [2] and can be downloaded from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). The dataset A (Accession number GDS 2206) was generated by using cDNA microarray based on 28 septal myocardial samples. It was composed of 13 DCM hearts at the time of transplantation and 15 nonfailing (NF) donor hearts. The generation of the dataset B (Accession number GDS 2205) was based on the oligonucleotide microarray study consisting of 12 independent subendocardial left ventricular samples: 5 and 7 samples were obtained from NF donors and DCM hearts respectively.

TABLE 1 THE CHARACTERISTICS OF DATASETS A AND B

Dataset	Platform	Tissues		Sex		Age (years)	
				Male	Female	Youngest	Oldest
A	Human Unigene3.1 cDNA Array 37.5K v1.0	Septum	DCM	8	5	14	60
			NF	11	4	43	58
B	Affymetrix GeneChip Human Genome U133 Array Set HG-U133A	Left ventricle	DCM	4	3	14	54
			NF	1	4	38	62

Both datasets were available in log-scale. Probe sets with missing values in more than 50% of their transcripts in either group were excluded in this study. The description of data preprocessing procedures used can be found in [2], [5].

B. Prediction models

A total of three statistical and machine learning models: Naive Bayes (NB), Support Vector Machines (SVM), and *K**, were employed to evaluate prediction performance of a set of genes or each individual gene.

NB is a simple probabilistic classifier based on Bayesian theorem with naïve independence assumptions. Despite this over-simplified assumption, NB is surprisingly successful in many practical applications [9]. NB-based classification combines the naïve Bayes probability model with the maximum a posterior decision rule. Let C be a dependent class variable with k classes $(C = c_1, c_2, \mathrm{K}, c_k)$ conditional on n feature variables $F_1, F_2, \mathrm{K}, F_n$. An NB-based classification model can be expressed as follows:

$$classify(f_1, f_{2,\mathrm{K}}, f_n) = \max p(c_j) \prod_{i=1}^{n} p(f_i \mid c_j), j \in [1,k] \quad (1)$$

where $f_1, f_{2,} \mathrm{K}, f_n$ are the value of features $F_1, F_2, \mathrm{K}, F_n$. $p(c)$ and $p(f \mid c)$ represent *prior* probability and likelihood respectively.

Based on solid theoretical foundation, SVM has demonstrated several important and unique features and has been applied to pattern recognition problems in a number of areas [10]. The main idea behind the SVM is to construct an *N*-dimensional hyperplanes as the decision-making surface that optimally classify samples into their respective categories. Conventional SVM-based classification involves the following steps: (a) training process; (b) identifying a set of support vectors; (c) computing the decision value, *sv*, for each test case using Equation (2); and (d) predicting the class of the test sample using a sign function.

$$sv = \sum_{i=1}^{k} \alpha_i y_i K(x, x_i) + b \quad (2)$$

Where k is the number of support vectors, x is the vector of a test case, x_i is i^{th} support vector, y_i is the output of i^{th} support vector, α_i is the associated Lagrange multiplier, and $K(x, x_i)$ is the kernel function.

*K** is a relatively simple instance-based classifier [11]. The class of a test case, *t*, is determined by finding training instance(s) that most similar to it. *K** uses an entropy-based distance measure to calculate the similarity between two samples. It computes the probability of the test case, *t*, belonging to class c_i using the following formula:

$$p(c_i \mid t) = \sum_{x_j \in c_i} p(x_j \mid t) \quad (3)$$

where x_j is the j^{th} member of class c_i. The class with the highest probability is determined as the classification of the test case, *t*.

C. Evaluation Methods

The quality assessment of a prediction model is generally based on the extent to which the correct category labels have been assigned. From the medical point of view, it is important not only to check how many test cases have been correctly identified, but also to examine how well a model can classify an unseen test case as not belonging to a particular case. Thus, this paper adopts three statistical measures: *accuracy* (AC), *sensitivity*(Se), and *specificity* (Sp), to evaluate classifiers.

Given the limited number of available training samples, a leave-one-out cross-validation is adopted. For a dataset with n samples, n experiments are carried out. For each experiment, a single sample is selected as the test case with the remaining samples used for training. The true error is estimated as the average error rate on test samples. To

further assess prediction performance, we also evaluated the classification models using a set of independent samples that is separated from the training set used to build it.

D. Implementation protocols

All three classification models were implemented within the framework provided by the *Weka* package [12]. Unless indicated otherwise, the models reported here using the following learning setting. The training of SVM is based on sequential minimal optimization algorithm developed by Platt [13], which breaks the large quadratic programming into a series of smallest possible quadratic programming. A polynomial kernel with exponent set to 1 was used. The parameter *C* representing the degree of tolerance was set to 1.0. The parameter *globalBlend* for K* classifier was equal to 50.

III. Results

A. Significance analysis

To establish the number of significantly differentially expressed genes, we performed *t*-test with Bonferroni correction. For dataset A, we found that more than 200 genes having an adjusted *p*-value less than 0.01 (raw p-value less than 4.61e-07) and close to 350 genes having an adjusted *p*-value less than 0.05 (raw *p*-value less than 2.3e-06). Examples of expression profiles of up-regulated and down-regulated genes are illustrated in Figs. 1 and 2.

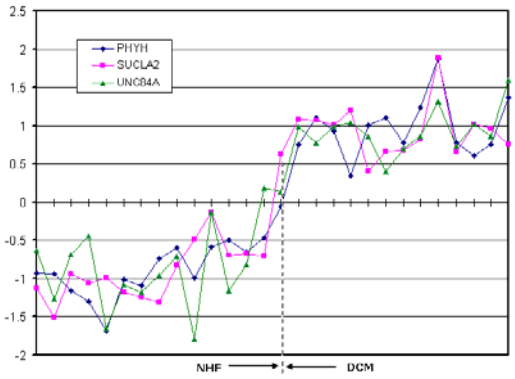

Fig. 1. Examples of expression profiles of up-regulated genes

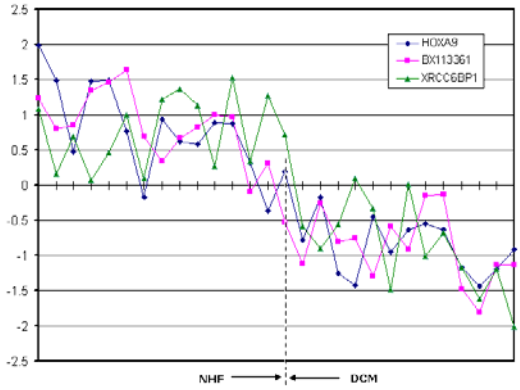

Fig. 2 Examples of expression profiles of down-regulated genes

Interestingly, many previously found to be biomarker genes of heart failure [2] do not show differentially expressed patterns in this dataset after Bonferroni correction. Examples include chemokine ligand 2 (CCL2), transcripts encoding for sarcomer structure proteins (MYH10), and the procollagen C-endopeptidase enhancer 2 (PCOLCE2). This is partly due to the stringent nature of Bonferroni correction technique.

We performed similar statistical analysis on Dataset B. Surprisingly we found that the set of differentially expressed genes in these two datasets had only a few in common. There are about 110 and 2580 genes, which have a raw *p* value below 0.001 in Datasets A and B respectively. Only the following 20 genes appeared in both sets: BACH2, C14orf133, C16orf45, CHST5, DHRS12, HMGN2, ISOC1, KIAA0195, KIAA0774, KIDINS220, MYH10, NCOA4, ODC1, RNF5, SEC31A, SGCE, SLC30A1, SSPN, TERF1, and YEATS2.

B. Predictive power of differentially-expressed genes

To assess predictive performance of each individual gene, we built the classifiers with one single variable. We firstly ranked the genes in terms of their p-values. Tables 2 and 3 show the prediction performance of top 10 genes of three classifiers for Datasets A and B respectively.

TABLE 2 PREDICTION PERFORMANCE OF THREE CLASSIFIERS WITH TOP 10 INDIVIDUAL FEATURES FOR DATASET A

Gene	NB	SVM	K*
BX089301	100%	100%	100%
PHYH	100%	100%	100%
BX094369	92.9%	92.9%	92.9%
LOC387890	96.4%	96.4%	96.4%
BX116905	96.4%	96.4%	96.4%
SUCLA2	96.4%	96.4%	96.4%
PSMD13	96.4%	96.4%	96.4%
UNC84A	100%	92.9%	96.4%
CSDE1	96.4%	96.4%	96.4%
PAIP2	100%	100%	100%

TABLE 3 PREDICTION PERFORMANCE OF THREE CLASSIFIERS WITH TOP 10 INDIVIDUAL FEATURES FOR DATASET B

Gene	NB	SVM	K*
TRMT5	100%	100%	100%
C16orf45	100%	100%	100%
CBFB	100%	91.7%	100%
H2AFZ	91.7%	100%	100%
AL133215	100%	100%	100%
SUV420H1	100%	75%	100%
ODC1	100%	83.3%	100%
SMC4	100%	83.3%	100%
IDH2	100%	91.7%	100%
C20orf149	100%	91.7%	100%

A closer examination of the results presented in Tables 2 and 3 reveals that:

1. Most classification with top 10 genes can achieve high classification accuracy (>90%). For example, For Dataset A, classification with gene PHYH and BX089301 obtained 100% accuracy using all three classifiers. 96.4% accuracy was achieved when using NB, SVM and K* with gene SUCLA2. For Dataset B, classification with TRMT5, C16orf45, and AL133215 with three classifiers can achieve 100% accuracy.
2. For Dataset A, three classification models exhibit the similar behaviour. The obtained AC values are between 92.9% and 100%. However, for Dataset B, there is a slightly different story with K* having consistent AC value (100%) and SVM demonstrating a relatively big variation of accuracy that ranges from 75% to 100%.
3. The ranking of each individual gene may not completely reflect its predictive power. To investigate the impact of gene ranking, we ranked genes with correlation coefficient-based ranking criterion [14], i.e. each gene, g_i, was ranked in terms of its capacity to distinguish between HF and DCM, which was defined as follows:

$$S_i = \frac{\mu_{NF}(g_i) - \mu_{DCM}(g_i)}{\sigma_{NF}(g_i) + \sigma_{DCM}(g_i)} \quad (4)$$

where $\mu_{NF}(g_i)$ and $\sigma_{NF}(g_i)$ be the mean value and the standard deviation of g_i in Class NF, $\mu_{DCM}(g_i)$ and $\sigma_{DCM}(g_i)$ be the mean value and the standard deviation of g_i for Class DCM. Tables 4 and 5 show the prediction results of top 10 genes. In comparison to the results presented in Tables 2 and 3, no significant difference was observed between these two ranking experiments.

TABLE 4 PREDICTION PERFORMANCE OF THREE CLASSIFIERS WITH TOP 10 GENES BASED ON CORRELATION COEFFICIENT RANKING CRITERION FOR DATASET A

Gene	NB	SVM	K*
BX089301	100%	100%	100%
PHYH	96.4%	100%	100%
BX116905	96.4%	96.4%	96.4%
BX094369	92.9%	92.9%	92.9%
PSMD13	96.4%	96.4%	96.4%
DLAT	92.9%	92.9%	92.9%
UNC84A	100%	92.9%	96.4%
LOC387890	96.5%	96.5%	96.5%
SUCLA2	96.5%	96.5%	96.5%
CSDE1	96.5%	96.5%	96.5%

TABLE 5 PREDICTION PERFORMANCE OF THREE CLASSIFIERS WITH TOP 10 GENES BASED ON CORRELATION COEFFICIENT RANKING CRITERION FOR DATASET B.

Gene	NB	SVM	K*
TRMT5	100%	100%	100%
C16orf45	100%	100%	100%
IDH2	100%	91.7%	100%
AL133215	100%	91.7%	100%
CBFB	100%	91.7%	100%
APOBEC2	83.3%	100%	100%
H2AFZ	91.7%	100%	100%
YEATS2	91.7%	91.7%	91.7%
SUV420H1	100%	75%	100%
KIAA0195	100%	91.7%	100%

C. Many gene sets can achieve a very high classification accuracy

We found that there were many sets of genes which can achieve 100% classification accuracy. We ranked the genes in terms of their adjusted p-values. For Dataset A, three prediction models using all genes with an adjusted *p*-value greater than 0.01 as input can achieve 100% prediction performance. The same 100% accuracy was obtained when using top 5, top10, top 50, top100, top 200 genes.

We then randomly selected 5 differentially-expressed genes as input to three classifiers. Once again we found that we can obtain very high classification performance. Using DCTN6, SCP2, MRFAP1L1, MRPL1, and C20orf29, for example, we obtained 100% classification accuracy. More than 96% classification accuracy was achieved with genes ATP6AP2, MFN2, ABI2, DAP3, and GSTZ1.

Similar observations were made when performing the analysis of Dataset B. Top 100 genes with the lowest p-values were selected and 100% classification accuracy was achieved for all three classifiers using top 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 genes.

The results presented in this section highlighted that it may not be sufficient to just examine a certain number of top genes for the identification of HF associated genes. Such main candidates should be chosen from the much longer list of genes. The similar observation was made by Ein-Dor *et al.* [6] when they studied signature genes in breast cancer.

D. Validation based on Independent dataset

To further assess predictive performance of each classification model, we evaluated them using an independent test dataset instead of leave-one-out cross validation. We firstly found a set of 35 genes which are significantly expressed (based on raw *p*-value < 0.01) in both Datasets A and B. Then we constructed 4 sets of training and testing samples which include 5, 10, 20 and 35 of these genes respectively. We used their expression values in one dataset for training while the prediction was based on their expression value in another dataset. The results were shown in Tables 6 to 13.

TABLE 6 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 5 GENES. THE EXPRESSION VALUES IN DATASET A WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET B. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*E. Sp* (%)
NB	100	100	100	100	100	100	100
SVM	100	100	100	100	100	100	100
K*	100	100	100	100	100	100	100

TABLE 7 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 10 GENES. THE EXPRESSION VALUES IN DATASET A WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET B. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*F. Sp* (%)
NB	100	100	100	100	100	100	100
SVM	100	100	100	100	100	100	100
K*	91.7	83.3	100	85.7	100	85.7	100

TABLE 8 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 20 GENES. THE EXPRESSION VALUES IN DATASET A WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET B. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES. THE PARAMETER OF GLOBALBLEND OF K* WAS SET TO 10

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*G. Sp* (%)
NB	83.3	71.4	100	71.4	100	71.4	100
SVM	91.7	83.3	100	85.7	100	85.7	100
K*	83.3	71.4	100	71.4	100	71.4	100

TABLE 9 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 35 GENES. THE EXPRESSION VALUES IN DATASET A WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET B. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*H. Sp* (%)
NB	84.6	75.0	100	71.4	100	71.4	100
SVM	92.3	85.7	100	85.7	100	85.7	100
K*	91.7	83.3	100	85.7	100	85.7	100

TABLE 10 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 5 GENES. THE EXPRESSION VALUES IN DATASET B WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET A. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES. A POLYNOMIAL KERNEL OF THIRD ORDER WAS USED IN SVM

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*I. Sp* (%)
NB	100	100	100	100	100	100	100
SVM	96.4	100	93.3	100	92.9	100	93.3
K*	100	100	100	100	100	100	100

TABLE 11 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 10 GENES. THE EXPRESSION VALUES IN DATASET B WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET A. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*J. Sp* (%)
NB	92.9	93.3	93.3	92.3	92.3	92.3	93.3
SVM	100	100	100	100	100	100	100
K*	89.3	100	80.0	100	81.3	100	80.0

TABLE 12 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 20 GENES. THE EXPRESSION VALUES IN DATASET B WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET A. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES. THE PARAMETER OF GLOBALBLEND OF K* WAS SET TO 50. A POLYNOMIAL KERNEL OF THIRD ORDER WAS USED IN SVM

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*K. Sp* (%)
NB	85.7	92.3	80.0	92.3	80.0	92.3	80.0
SVM	92.9	93.3	93.3	92.3	92.3	92.3	93.3
K*	60.7	100	26.7	100	54.2	100	54.2

TABLE 13 PREDICTION RESULTS FOR THREE CLASSIFIERS WITH 35 GENES. THE EXPRESSION VALUES IN DATASET B WAS USED FOR TRAINING WHILE THE PREDICTION WAS BASED ON THEIR EXPRESSION VALUES IN DATASET A. NB USES SUPERVISED DISCRETIZATION TO CONVERT NUMERIC ATTRIBUTES TO NOMINAL ONES

Model	*Ac* (%)	NF			DCM		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*L. Sp* (%)
NB	82.1	91.7	73.3	92.3	75.0	92.3	73.3
SVM	89.3	87.5	93.3	84.6	91.7	84.6	93.3
K*	57.1	100	20.0	100	52.0	100	20.0

As can be seen from the results presented in these tables, training with Dataset B and testing with Dataset A generally produced the worse performance comparing with that of training using Dataset A and testing using Dataset B. This is because that Dataset B was composed of fewer samples. Such a difference was particularly significant when the number of genes was getting larger and instance-based classifier (K*) was used.

Another observation was that using the 5 genes can achieve the best results while continual increase of genes doesn't contribute to the improvement of prediction performance. In fact, in most of cases, the poor results were achieved when using all 35 genes.

IV. Discussion and Conclusions

Based on the analysis of two publicly available HF microarray data, this paper evaluated the sets of HF related genes derived from computational analysis of gene expression profiles. While we found that many genes were highly differentially expressed in one dataset, there were only a few common genes differentially expressed in both datasets. More importantly, we found that many sets of genes can achieve 100% prediction accuracy. Classification with several individual genes such as PHYH and TRMT5 can also obtain very high accuracy. This may suggest that it should not be enough to just focus a certain number of top genes for the identification of potential targets for therapy. Such candidates should be chosen from the much longer list of genes [6]. Based on validation using independent datasets, we found that classification with, for example, 5 genes, i.e. SEC31A, PIK3CA, SSPN, ODC1, and KIAA0195 can achieve 100% prediction accuracy.

Recent years have seen growing interest in studying cardiovascular biomarkers. Morrow and de Lemos [15] have set out three criteria for the appraisal of novel biomarkers: (a) Can the clinician measure it? (b) does it provide new information that is not already available? And (c) will the identified biomarkers help the clinician make decision? Eugene [16] argued that only relatively few of the biomarkers previously identified can satisfy all three criteria. However, they may provide important relevant information such as the pathogenesis of heart failure. He then divided a list of HF biomarkers into six categories: (a)Inflammation; (b) Oxidative stress; (c) Extracellular-matrix remodeling; (d) Neurohormones; (e) Myocyte injury; and (f) Myocyte stress. Comprehensive analysis of expression profiles of these biomarkers would be an important part of our future work.

Currently we only analyzed two HF expression data with three classification models. It is worth repeating the experiments using more microarray datasets associated with HF. The behaviour of other types of machine learning and statistical models such as neural network based classifiers also deserves further investigation.

References

[1] Rosamond *et al.*, "Heart disease and stroke statistics – 2008 update," *Circulation*, 117: e25-e146, 2008.

[2] A. Barth, R. Kuner, A. Buness, M. Ruschhaupt, S. Merk, L. Zwermann, S. Kääb, E. Kreuzer, G. Steinbeck, U. Mansmann, "Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies," *J Am Coll Cardiol.*, 48(8), pp.1618-1620, 2006.

[3] F. Tan, C.S. Moravec, J. Li, C. Apperson-Hansen, P.M. McCarthy, J.B. Young, M. Bond, "The gene expression fingerprint of human heart failure," *Proc. Natl. Acad. Sci. USA* (PNAS), 99, pp.11387-11392, 2002.

[4] F. Wittchen, *et al.*, "Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets," *J Mol Med*., 85 (3) pp.257-271, 2007.

[5] A. Camargo and F. Azuaje, "Identification of dilated cardiomyopathy signature genes through gene expression and network data integration," *Genomics*, in press, 2008.

[6] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?," *Bioinfromatics*, vol. 21, no. 2, pp.171-178, 2006.

[7] I.S. Lossos, D.K. Czerwinski, A.A. Alizadeh, M.A. Wechser, R. Tibshirani, D. Botstein, and R. Levy, "Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes," *N. Engl. J. Med.*, **350**, pp.1828–1837, 2004.

[8] G.L. Miklos and R. Maleszka, "Microarray reality checks in the context of a complex disease," *Nat. Biotechnol.*, **22**, pp.615–621, 2004.

[9] R. Irina, "An empirical study of the naive Bayes classifier," *in the Proc. Of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligenc*e, 2001.

[10] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2, pp.121-167, 1998.

[11] J. G. Cleary and L.E. Trigg, "K*. An Instance- based learner using an entropic distance measure," *In the Proc. of the 12th International Conference on Machine learning*, 108-114, 1995.

[12] I.H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[13] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning," B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press, 1998.

[14] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeck, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander," "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*,, 286, pp.531-537, 1999.

[15] D.A. Morrow and J. A. de Lemos, "Benchmarks for the assessment of novel cardiovascular biomarkers," *Circulation*, 115, pp. 949-952, 2007.

[16] E. Braunwald, "Biomarkers in heart failure," *The New England Journal of Medicine*, 358, pp.2148-2159, 2008.